Modification of the Support Vector Machine Method for Classification and Detection of Anomalies in Image Processing Problems

Volodymyr Kobziev, Kateryna Goryshnia
dept. of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua, kateryna.horishnia@nure.ua

Sergiy Yakovlev
National Aerospace University "Kharkiv Aviation Institute", Kharkiv,Ukraine
Lodz University of Technology, Lodz, Poland
sergiy.yakovlev@p.lodz.pl

Модифікації Методу Опорних Векторів для Класифікації та Виявлення Аномалій у Задачах Обробки Зображень

Володимир Кобзєв, Катерина Горішня
Кафедра Програмної інженерії
Харківський Національний університет радіоелектроніки
Харків, Україна
volodymyr.kobziev@nure.ua, kateryna.horishnia@nure.ua

Сергій Яковлев
Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна
Лодзький Університет технологій, Лодзь, Польща
sergiy.yakovlev@p.lodz.pl

***Abstract*—Various aspects using modifications of the support vector machine method for variants of linear and non-linear separation of classification objects with the greatest distance between them are considered.**

***Keywords*—classification; anomaly detection; support vectors machine method; maximum gap; linear and non-linear separation**

I. Introduction

Modern analysis of data accumulated and supplemented over time in many applied fields involves the study of various images (space, medical, topographic, etc.) with the aim of identifying existing patterns in them [1]. For Data Mining tasks used in such analysis, methods of classification and anomaly detection are important both for permanent data [1] and for time series [2].

In the majority of applied studies, the division of the existing population of some objects into two parts according to a certain ratio of their spatial features is considered, as well as the detection of objects that differ significantly in the set of values of their characteristics from other elements of the population. The support vector method uses prior training and definition of object classification rules and makes it possible to solve these problems in a relationship.

II. Main matherial

The method of support vectors machine [3] for the training population, which is given by a set of vectors $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ in the hyperspace R^k, determines such values of the coordinates of the normal vector $\boldsymbol{\omega}$ and the constant β, which make it possible to present the equation of the separating hyperplane $\boldsymbol{H}$ of two classes in the form:

$$\boldsymbol{\omega}^{\boldsymbol{T}} \bullet \boldsymbol{\xi} + \beta = 0. \qquad (1)$$

The displacement parameter β coincides with the coordinate of the point of intersection of the hyperplane with the ξ axis, the vector $\boldsymbol{\omega}$ is the vector of weights, which are chosen so that equality (1) is fulfilled. Such a hyperplane is at an biggest equidistant distance from the nearest representatives of two different classes in the training set, which play the role of support vectors.

The distance from the origin to the hyperplane $\boldsymbol{H}$ is equal to $(|\beta|)/(\|\boldsymbol{\omega}\|)$, where the denominator means the Euclidean norm (length) of the vector w. The two hyperplanes $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ containing the mentioned nearest objects of different classes are parallel to the hyperplane $\boldsymbol{H}$. The distance between $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ (the boundary width or gap separating the two classes) is $\rho = 2 / \|\boldsymbol{\omega}\|$. If we remember that $\|\boldsymbol{\omega}\| = \sqrt{(\boldsymbol{w}^{\boldsymbol{T}} \bullet \boldsymbol{\omega})}$, then the desire to maximize the width of the gap ρ as much as possible leads to the need to minimize the quadratic function under linear constraints. The quadratic optimization problem is a well-studied mathematical optimization problem, for the solution of which specialized methods of quadratic programming are proposed to implement the support vector method.

When using this classification method, new objects that do not cross the hyperplane $\boldsymbol{H}_1$ in the direction of $\boldsymbol{H}_2$ in space are assigned to the first class. Objects that do not cross the hyperplane $\boldsymbol{H}_2$ in the direction of $\boldsymbol{H}_1$ in space belong to the second class, respectively. At the same time, objects whose coordinates in space fall into the gap between hyperplanes $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ (do not belong to either of the two classes) can be considered anomalous.

The described approach is suitable for problems of very large dimensions, which allow a linear division of objects into classes, which in general is not performed. If this condition is fulfilled, then preference is still given to the solution that better separates the bulk of the data, ignoring a small number of unusual noise objects.

If the training set is not linearly separable, then, of course, when constructing a wide separation gap (boundary), a number of errors may occur (individual points - anomalous instances - may lie both inside the gap and on the wrong side). In such cases, we speak of a classification problem with a soft gap. For each incorrectly classified instance, a penalty will be imposed, depending on how much the conditions imposed on the gap are violated.

Under such conditions, dummy variables $\zeta i \geq 0$ are introduced into the optimization problem.

$$\boldsymbol{\xi}' = (\xi_1+\zeta_1 , \xi_2+\zeta_2 , \ \dots , \xi_n+\zeta_n). \quad (2)$$

Taking the dummy value $\zeta i > 0$, the width of the gap for the point xi can be made less than one, but at the same time you will have to pay a penalty C•ζi. The sum of values ζi determines the upper limit of the number of errors during training. The number of errors during training due to the width of the gap minimizes the method of support vectors with a soft gap (soft-margin SVM). At the same time, the solution of the optimization problem is a compromise: it establishes a balance between the width of the gap and the number of points that would have to be moved in order to ensure this width. Parameter C allows you to control the retraining: if it becomes large, it is undesirable to ignore the data due to the reduction of the geometric gap; if its value is small, then with the help of dummy variables it is not difficult to take into account some points and get the size of the gap that models the bulk of the data.

As a rule, the support vectors make up a small part of the training set. The training time in the support vector method is mainly determined by the time of solving the corresponding quadratic programming problem, therefore the theoretical and empirical complexity of this stage depends on the method of solving this problem. The dependence of the time complexity of the standard solution of the quadratic programming problem on the amount of training data is considered to be cubic. The nonlinear empirical complexity of traditional support vector algorithms makes it difficult and may make it impossible to apply them to large sets of training data. All the latest works on the support vector method are aimed at reducing this complexity, and most often by replacing the exact solution with an approximate one. It should be noted that only a set of points close to the desired separation limit is used at the training stage. This significantly speeds up the learning process compared to known types of neural networks.

One of the ways to solve the classification problem in cases where the data sets do not allow linear division is to map the data into a higher dimensional space followed by applying a linear classifier in this space. There are known examples when a linear classifier easily recognizes data when using a quadratic function or polar coordinates to map the data to a two-dimensional plane. The main idea is to map the original feature space into a feature space of a higher dimension, in which the training set turns out to be linearly separable. Of course, at the same time, it is desirable to preserve the relevant dimension of relations between data points so that the resulting classifier summarizes the original data.

If the problem does not have the property of linear resolution or the gap is small and the set of incorrectly classified points or points lying inside the gap will be large enough, then in the non-linear case this may turn out to be the main factor in reducing the slowdown of the support vector method at the stage of classification of new data. Also, in cases of a considerable number of incorrectly classified points, SVM modification using kernel functions is applied.

To significantly accelerate the solution of the general problem with the smallest classification errors, it is possible to propose the parallel execution of various SVM variants and the selection of the most adequate one based on their results, as used in the paper [4] for another tasks.

The described variants of the method of support vectors can be used in various application areas, in particular, where it is possible to divide the studied objects into two types, for example, one's own - someone else's, healthy - sick, etc.

REFERENCES

[1] Han, Jiawei. Data mining: concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed., Morgan Kaufmann Publishers is an imprint of Elsevier, 2012, 740p.

[2] L. Kirichenko, V Kobziev, Y Fedorenko. Data Mining methods for detection of collective anomalies in time series // Intern. sc.-pr. conf. "Application of information technologies in the preparation and operation of law enforcement forces" / Collection of theses of reports. – Kharkiv. – 2021. - p. 106.

[3] Steinwart, Ingo; and Christmann, Andreas; Support Vector Machines, Springer-Verlag, New York, 2008.

[4] S. Yakovlev, A. Khovrat, V. Kobziev. Using Parallelized Neural Networks to Detect Falsified Audio Information in Socially Oriented Systems, in: Proceedings of the International Scientific Conference "Information Technology and Implementation", IT&I'23, Kyiv, Ukraine, 2023, pp. 220–238. URL: https://ceur-ws.org/Vol-3624/Paper_19.pdf.